Iurii Kolomeitsev, Nurlan Shagadatov

Hybrid Latent Semantic Indexing

Machine Learning Project

Skolkovo Institute of Science and Technology, 2018

Contents

1 Introduction

In this project we took an idea of Hybrid SVD decomposition proposed by E. Frolov and I. Oseledets in [1] for recommender systems. This method allows to incorporate side information of objects and features in a generalized way that gives substantial increase in quality of analysis. We apply this approach in another field of data analysis, namely text mining by substituting standard SVD in Latent Semantic Indexing (LSI) method with Hybrid SVD and get a new method of text analysis called Hybrid LSI. We test Hybrid LSI method on document classification tasks, and see how we can benefit from it in comparison with using standard LSI, Relevant Documents Selecting Method (RDS), Local Relevancy Weighted (LRW) LSI and not using LSI at all. For testing we use 2 datasets: 20Newsgroups [2], Reuters-21578 [3] and Fasttext word embeddings [4]. We tune hyperparameters of methods: hyperparameters of a classifier, hyperparameters of Hybrid LSI, hyperparameters of RDS and LRW methods, the number of components and see in what settings Hybrid LSI increases the quality of classification.

2 Related Work

A great number of works propose improvements of the LSI method, especially for the text classification task. In [5] authors propose to use background, unlabeled text. In [6] authors propose the concept of local LSI which integrates the class information and performs separate SVD on the local region of each topic. In [7] authors propose a modification of local LSI method called Local Relevancy Weighted LSI which improved Hull's results.

An interesting model closely related to the LSI is Probabilistic Latent Semantic Indexing (PLSI) [8]. While LSI stems from linear algebra, PLSI has a statistical foundation and defines a generative model of the data. This approach also results in the matrix decomposition, however, obtained by maximizing likelihood of the data.

3 Problem formulation

3.1 Document classification

Document classification task is to assign a document to one or more classes or categories. Firstly, we should extract features out of documents and construct a document-term matrix of word occurrences in documents using, for example, bag-of-words, term frequency (tf) or term frequency – inverse document frequency (tf-idf) representations. Tf-idf takes into account that some words can occur in many documents regardless of a document's class and gives these words smaller weights. That gives better quality, so further we will use only tf-idf scheme.

The resulting document-term matrix is very sparse and maybe noisy, so a good thing to do is to find it's low-rank approximation using SVD. This method is known as LSA/LSI and it is based on the principle that words that are used int the same contexts tend to have similar meanings. This method was originally developed for information retrieval tasks [9] and later became widely used in many other text analysis problems.

Nevertheless, standard LSI model does not take into account additional information such as relations between documents or terms. This brings us to the idea of using Hybrid SVD to incorporate this information. As term similarity matrix we can take use any word embeddings model, for example, Word2vec or Fasttext, and compute pairwise cosine similarities between vectors representing words. Document similarity matrix depends on a particular task. For example, if we know the domains (or authors) of documents then it could be a zero-one matrix with ones, if the 2 documents are from the same domain and zeros – from different.

3.2 Hybrid SVD

The latent factor model of SVD can be viewed as an eigendecomposition of a scaled document-based or term-based cosine similarity matrix. In a document-based case it solves an eigendecomposition problem for the following matrix:

$$A = RR^T = DCD, \tag{1}$$

where $R \in \mathbb{R}^{D\times T}$ – document-term matrix, $D \in \mathbb{R}^{D\times D}$ – diagonal scaling matrix with elements $d_{ii} = \|r_i\|$ – Euclidian norm of a vector r_i that is an i-th row of the matrix R. Matrix C consists of cosine similarities between documents – rows of R:

$$c_{ij} = \cos(i, j) \sim r_i^T r_j. \tag{2}$$

It is easy to see that any cross-term relations are simply ignored by SVD as the cosine similarity takes only term-to-term co-occurrence into account, so the contribution of a particular term into the final similarity score c_{ij} is counted only if the term is present in both documents i and j.

Thus, to take synonyms into account we can introduce term similarity matrix $S \in \mathbb{R}^{T\times T}$:

$$\text{sim}(i, j) \sim r_i^T S r_j. \tag{3}$$

By analogy we introduce document similarity matrix $K \in \mathbb{R}^{D\times D}$ to incorporate document-related information.

Then we can rewrite the eigendecomposition problem in the form of a system of equations:

$$\begin{cases} RSR^T = U\Sigma^2U^T, \\ R^TKR = V\Sigma^2V^T, \end{cases} \tag{4}$$

where $U \in \mathbb{R}^{D\times r}$, $V \in \mathbb{R}^{T\times r}$ – embedding of documents and terms onto a latent feature space, $\Sigma \in \mathbb{R}^{r\times r}$ – diagonal with singular values on diagonal.

The solution of this system is a normal SVD of matrix $K^{\frac{1}{2}}RS^{\frac{1}{2}}$:

$$\hat{R} = K^{\frac{1}{2}}RS^{\frac{1}{2}} = \hat{U}\Sigma\hat{V}^T, \tag{5}$$

where $\hat{U} = K^{\frac{1}{2}}U$, $\hat{V} = S^{\frac{1}{2}}V$ – matrices with orthonormal columns.

Finding the square root of a matrix is a computationally hard problem. Therefore, let's restrict matrices K and S to be symmetric and positive definite, so they should have diagonal dominance:

$$K = I + \alpha K', \quad S = I + \beta S', \tag{6}$$

where K', S' — original zero-diagonal symmetric similarity matrices, $0 \leq \alpha < 1$ and $0 \leq \beta < 1$ – tunable parameters of the HSVD model. As K and S are symmetric and positive definite, we can use Cholesky decomposition to find thier square roots:

$$K = L_kL_k^T, \quad S = L_sL_s^T \tag{7}$$

Matrices K and S can be sparse, so in such cases we can use sparse Cholesky decomposition or incomplete Cholesky decomposition.

4 Local LSI

The methods which are applied to the whole training set ignore class discrimination while only concentrating on representation.Some local LSI methods have been proposed to improve the classification by utilizing class discrimination information. It performs a separate SVD on the local region of each topic. Compared with global LSI, this method utilizes the class information effectively, so it improves the performance of global LSI greatly. We have considered two different approaches to define local spaces and compared their performances with Hybrid LSI.

4.1 Relevant Documents Selecting Method (RDS)

RDS defines the local region for a topic as the relevant documents only. It is the simplest method but the local region contains no discrimination information, so it is very limited to improve the classification performance.On the other hand, the frequency of topic occurrence varies greatly from topic to topic. As a result, method suffers from class imbalance.

4.2 Local Relevancy Weighted LSI

The 0/1 weighting method is a simple but crude way to generate local region. It assumes that the selected documents are equally important in the SVD computation. However, it is obvious that each document plays a different role in the local semantic space and the more relevant documents should contribute more to the local semantic space, and vice versa. LRW-LSI deals with this problem by weighting documents according to their relevancy score of correspondence for each class.

As a result, the training process of LRW-LSI contains the following six steps. At the first step, the initial classifier IC of topic c is used to assign initial relevancy score rs to each training document. Then at step two, each training document is weighted according to equation

$$d_i = d_i * f(rs), \quad \text{where} \quad f(rs) = \frac{1}{1 + e^{-a(rs_i + b)}}$$

The weighting function f is a Sigmoid function which has two parameters a and b to shape the curve. At step three, the top $m\gamma$ documents are selected to generate the local term-by-document matrix of the topic c (m is the number of documents of training sample corresponding to class c). Then at step four, a truncated SVD is performed to generate the local semantic space. At step five, all other weighted training documents are folded into the new space. Eventually, all training documents in local LSI vector are used to train a real classifier RC of topic c.

5 Experiments

The Hybrid LSI approach was tested on 2 datasets: 20 Newsgroups, Reuters-21578. All datasets were represented as tf-idf matrices with fixed vocabulary, based on 8000 top words ordered by term frequency across the corpus. The quality of classification was measured by accuracy, precision, recall and f1 scores averaged using weighted, micro and macro schemes (plots are provided with only weighted f1 score as the relative positions of methods for other scores are almost the same). Below is a table with main specifications of the datasets:

dataset	number of documents	number of classes	average documents length	initial sparsity, %	sparsity, %
20 Newsgroups	18846	20	181.6	0.066	0.858
Reuters-21578	10739	60	127.61	0.1625	0.5938

Table 1: Datasets specifications

As a classifier in all tasks we took linear SVM with l2 penalty and squared hinge loss function. The multiclass classification is handled according to a one-vs-the-rest scheme and the scores in such cases are weighted averages of scores of each class. As a baseline solution we took the classification without dimensionality reduction – using full tf-idf matrix which had size: number of documents × vocabulary size.

5.1 20 Newsgroups

The 20 Newsgroups data set is a collection of 18846 newsgroup documents, partitioned nearly evenly across 20 different newsgroups. Train set consists of 11314 documents, test set consists of 7532 documents. The split between the train and test set is based upon a messages posted before and after a specific date.

5.2 Reuters-21578

Reuters-21578 is one of the most commonly used dataset for text classification, it was used in some of the most influential papers on the field, for example in [10]. The dataset consists of news articles that can be assigned to several classes which is a multi-class and multi-label problem. The collection originally consists of 21578 documents but a subset and split is traditionally used. We use Mod-Apte split and consider only those categories that have at least 10 documents. Finally, we get train set of size 7741, test set of size 2998 and 60 categories.

5.3 Hyperparameters tuning

We tuned LSI and Hybrid LSI hyperparameters using 5-fold cross validation on train set. For LSI method we tune hyperparameters C of SVM classifier for each component. For Hybrid LSI method we tune for each component we tune α and C. The results are shown on Figures 1, 2 and 3.

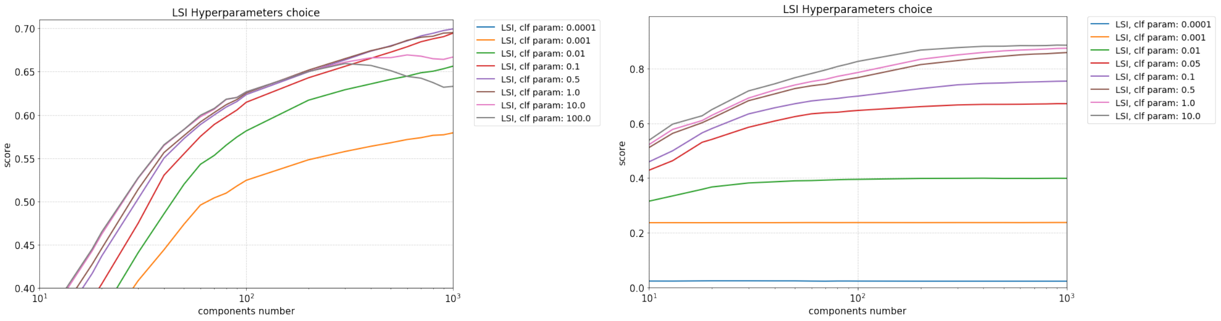

Figure 1: LSI hyperparameters tuning on 20Newsgroups (left) and Reuters-21578 (right)

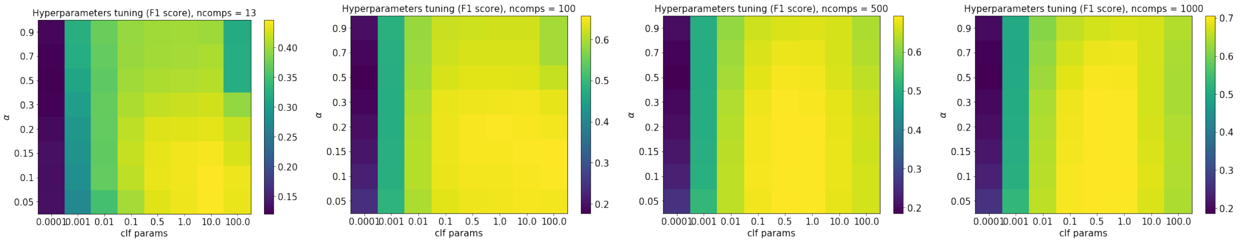

Figure 2: Hybrid LSI hyperparameters tuning on 20Newsgroups

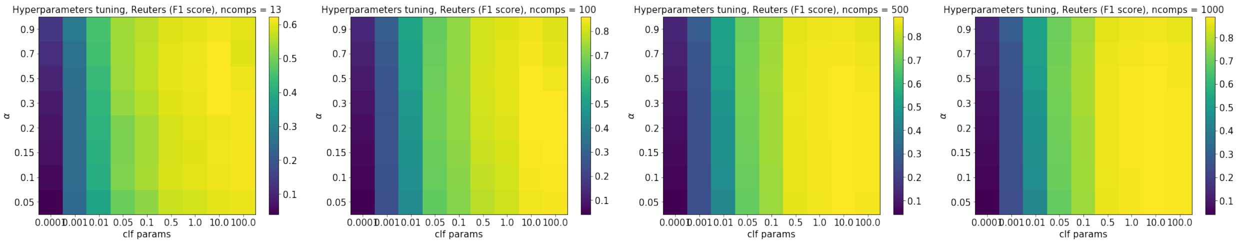

Figure 3: Hybrid LSI hyperparameters tuning on Reuters-21578

Local LSI methods: RDS and LRW were tuned using validation set (25% of train). For For both methods we have tuned parameter C of SVM Classifier. For LRW LSI we found optimal values of a, b and γ. The results of tuning RDS method are shown on Figure 4, Figure 5. The results of tuning LRW method are shown on Figure 6.

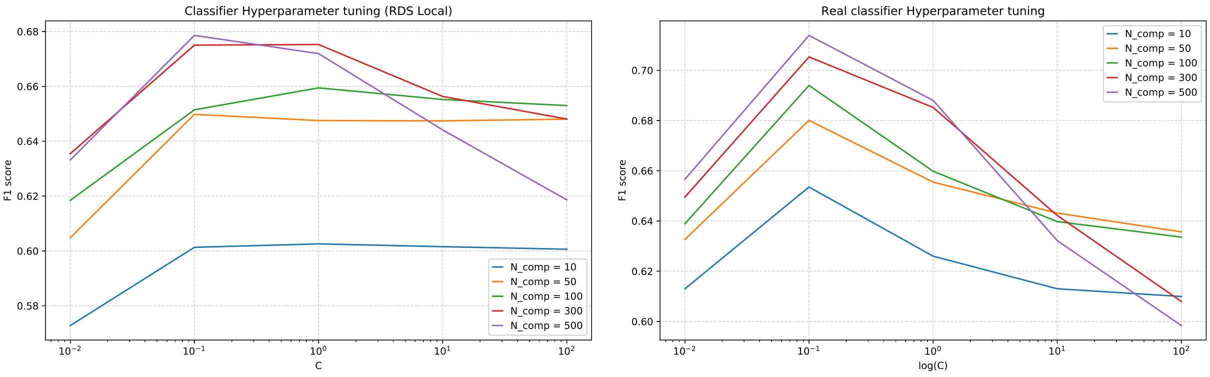

Figure 4: RDS hyperparameters tuning on 20Newsgroups

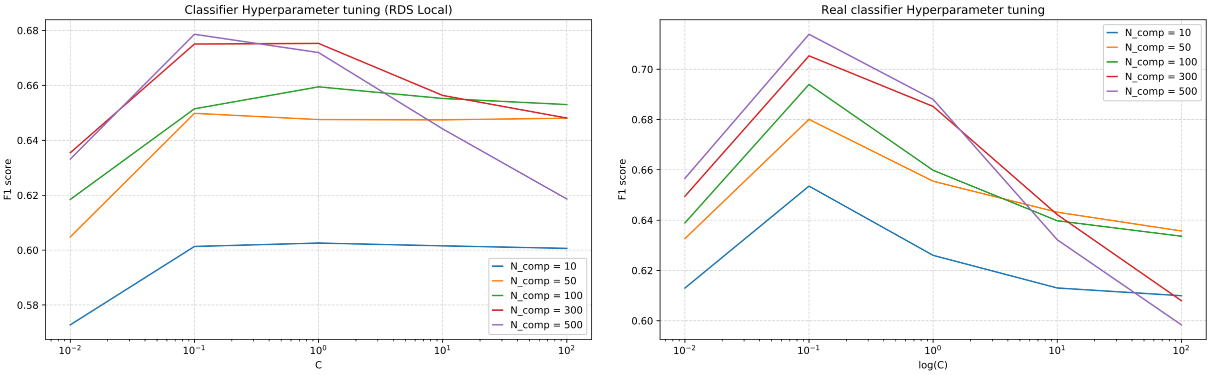

Figure 5: RDS hyperparameters tuning on Reuters-21578

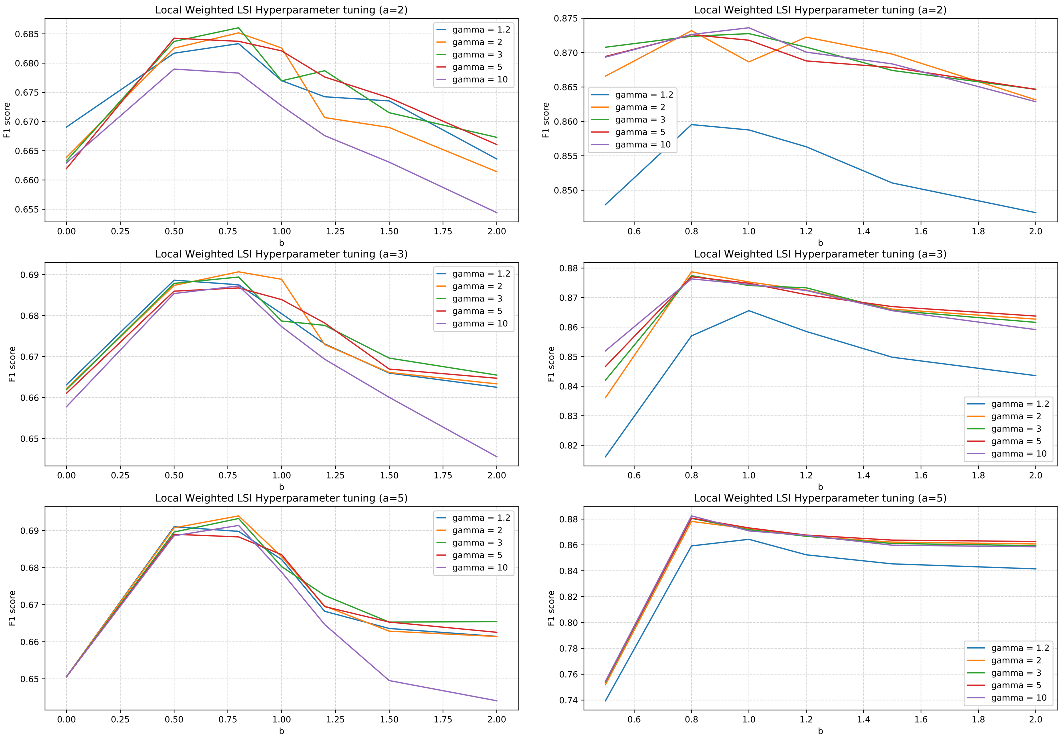

Figure 6: LRW hyperparameters tuning on 20Newsgroups (left) and Reuters-21578 (right)

5.3.1 Results

The final results with optimal hyperparameters are shown on Figure 7. In the case of the 20Newsgroups dataset, HybridLSI gives in average 5% increase of weighted-averaged F1 measure. In the Reuters-21578 dataset case the increase in quality is not as great as for 20Newsgroups, though the tendencies are the same: HybridLSI is consistently better than standard LSI.

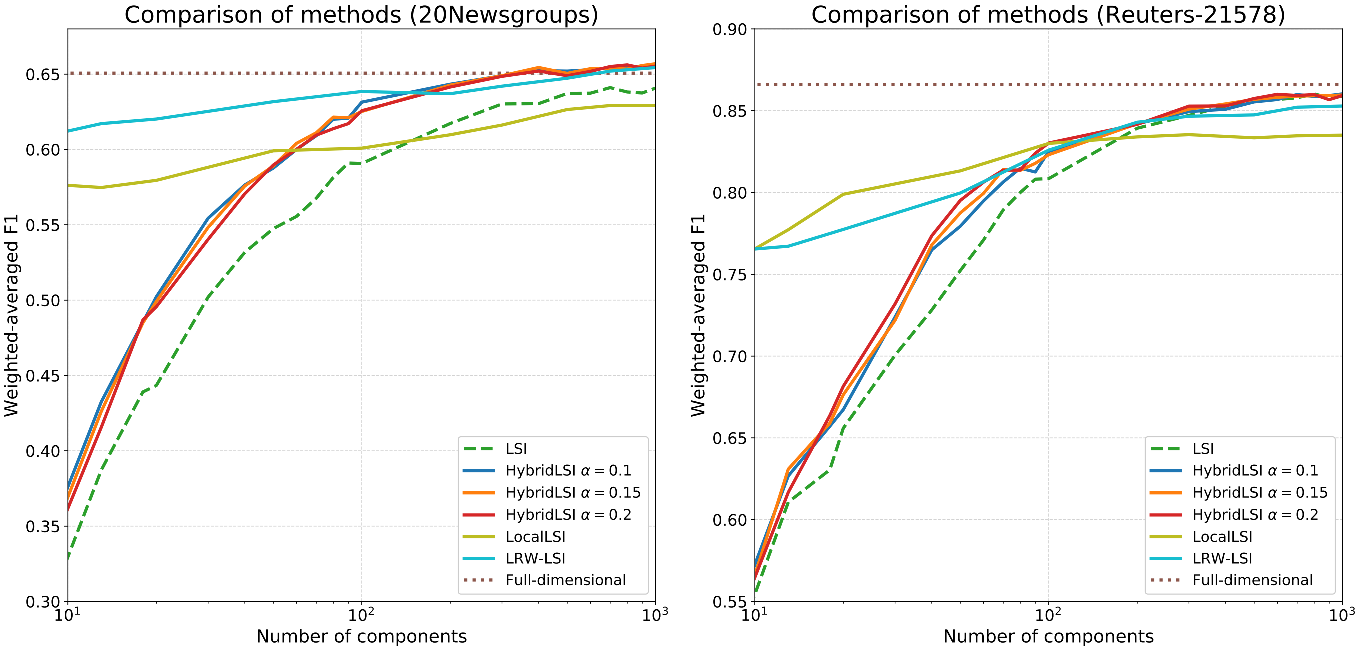

Figure 7: Comparison of different modifications of LSI on 20Newsgroups (left) and Reuters-21578 (right) datasets.

6 Conclusion

In this project we have successfully applied Hybrid SVD approach to the document classification task. We compared this method with standard LSI and two Local methods that improves LSI: Relevant Documents Selecting method and Local Relevancy Weighted LSI. These methods have been tested on the documents classification task for 2 datasets. We demonstrated that in all considered cases Hybrid LSI consistently outperforms standard LSI. In some cases it outperforms Local LSI methods. We observed that Local LSI techniques allows to achieve great quality even on low ranks.

7 Future work

The proposed method for text mining seems to be very promising, so we are planning to continue the research in this area. In particular, the following points could be investigated:

- explore different term similarity measures;
- develop approaches to the other text mining problems (e.g. clustering, textual similarity);
- combine Local LSI and Hybrid LSI together;
- end-to-end solution where S and K are part of optimization process.

8 Contribution

Iurii: implementing and testing LSI and Hybrid LSI method on 20Newsgroups, Reuters-21578 datasets, tuning hyperparameters of LSI and Hybrid LSI methods, making presentation, writing report.

Nurlan: implementing and testing RDS and LRW LSI methods on 20Newsgroups, Reuters-21578, tuning hyperparameters of RDS and LWR LSI, making presentation, writing report.

References

[1] E. Frolov and I. Oseledets. Hybridsvd: When collaborative information is not enough. *ArXiv e-prints*, February 2018.

[2] 20 newsgroups [dataset]. `http://qwone.com/~jason/20Newsgroups/`.

[3] Reuters-21578 [dataset]. `http://www.daviddlewis.com/resources/testcollections/reuters21578/`.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *CoRR*, abs/1607.04606, 2016.

[5] Sarah Zelikovitz and Haym Hirsh. Using lsi for text classification in the presence of background text. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 113–118, New York, NY, USA, 2001. ACM.

[6] David Hull. Improving text retrieval for the routing problem using latent semantic indexing. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 282–291, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[7] Tao Liu, Zheng Chen, Benyu Zhang, Wei-ying Ma, and Gongyi Wu. Improving text classification using local latent semantic indexing. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, ICDM '04, pages 162–169, Washington, DC, USA, 2004. IEEE Computer Society.

[8] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[9] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R.A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41*, pages 391–407, 1990.

[10] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML'98, pages 137–142, Berlin, Heidelberg, 1998. Springer-Verlag.